2.

Divergence and Probability in Taxonomy.

ISAAC GINSBURG

U. S. Bureau of Fisheries, Washington, D. C.

Taxonomists of past generations have generally been content with describing and establishing species based on one or but a very few specimens. The business of distinguishing species by this method is a comparatively easy matter. Using very few specimens as a basis of comparison, closely related species, in their large majority, appear to be sharply differentiated. In occasional instances a sharp distinction on the basis of even a few specimens proved troublesome, and such specimens were generally assumed to represent "varieties," "races," etc., of the same species.

This easy method proved to be inadequate, as it was bound to. Later investigators found that such distinctions all too frequently did not accord with their material. This is due to the fact that related species often approach closely or even intergrade in their differentiating characters. From a taxonomic point of view we do not know all we should about a species until we know its range and manner of variability, at least in the few crucial characters by which it is distinguished from closely related ones. This is, of course, true of races and other subdivisions of the species. Taxonomists come more and more to realize this and act accordingly. In dealing with mass data obtained in the study of variability, it is desirable to reduce them, when it is consistent to do so, to single figures, statistical constants, for convenience in comparison, discussion and interpretation. This paper considers one such class of constants, that concerned with the measure of divergence, which is of the utmost importance to taxonomists, as related to another class, that concerned with the expression of probability, which is often used in place of the first.

Probability in its numerical expression is often referred to as the "test of significance." As it is my hope that this paper will prove to be of interest to taxonomists to whom the latter term is not a household phrase, it may be well to consider briefly here its precise meaning.

When a taxonomist compares the likenesses of and the differences between two closely related populations—be they species, subspecies, races, etc.—he does not study the variability of the entire population, but his comparison is based on a relatively restricted number of specimens; in other words, on two samples drawn one each from the two populations. The degree of difference or divergence shown by the two samples determines his conclusion regarding the taxonomic rank of the two populations, whether they are to be considered as species or as belonging to a category of the next or second next lower rank. However, we know that different samples drawn even from the same variable population will generally not be the same, but, on the contrary, due solely to chance, will exhibit differences of greater or lesser degree. The question then comes up, is the difference shown by the two samples compared in taxonomic research a real population difference, or is it due to the fortuities of sampling? It may be assumed that due solely to chance it may happen sometimes that two

samples drawn even from the same population will exhibit a difference as great or greater than that between the two samples of the two populations compared, and in that case it would, of course, be inappropriate to separate the two samples taxonomically. An answer to the above question, in part, as discussed below, is given by certain mathematical formulae developed in connection with the theory of probability. By the use of such formulae —based on the difference between the averages of the two samples, the squares of the deviations of the individual specimens from the averages, and the number of specimens in the samples—it may be determined, for any one given character, how often such a difference, or a greater difference, is likely to be obtained at random, by mere chance, from two samples of the same population. If such an eventuality is likely to occur but rarely, say, two times or less per 100 trials, we may state that the difference is "significant" and that it may be concluded with comparative assurance that the two samples compared in our taxonomic research belong to two distinct populations.

The above is a bare outline of the "test of significance," but it is hoped that it presents its essential idea. The "test of significance" then results in a number that represents the numerical expression of probability, indicating the probable value of a difference determined in biologic research, or the probable reliability of the two samples compared, for the purpose of drawing pertinent conclusions. In the practical application of the formulae an arbitrary limit is postulated and a figure obtained as a result of the test of significance, which equals or is higher than the accepted limit, is taken to denote "significance." It should be noted in particular that this test merely establishes that a determined difference is "significant." It does not indicate definitely whether the difference is of specific, subspecific or racial magnitude. The taxonomic rank of the two populations compared is determinable definitely only by some appropriate measure of divergence.

It is very important and can not be too strongly emphasized that it is necessary to draw a sharp distinction of the fundamental difference between the two concepts, measure of divergence and expression of probability, from both a theoretical and a practical standpoint. This fundamental idea has been formulated by Fisher (1936, p. 59) as follows: "It must be stressed that the test of significance calculates a probability; it does not calculate a racial difference." Although it is, or should be, generally realized that a test of probability is not the same as a measure of divergence, yet, somehow, the two concepts become inextricably mixed in deliberation and discussion. Somehow or other there appears to be a lingering idea with some biologists that the greater the numerical value of the figures showing "significance" obtained by the use of current formulae that express probability, the greater the divergence between the pair of populations compared. Often this is true; but it is only a partial truth, and like all partial truths it is bound to lead us sadly astray. This confusion of concepts appears to be a stumbling block not only with biologists who are not given much to the employment of statistical formulae, but even with some who employ them extensively.

If two separate comparisons be made of two pairs of populations, and the test of significance have a much greater numerical value as between one pair of populations than between the other, it does not always mean that the former pair diverges to a greater extent; although in many cases a greater numerical value for the test of significance does coincide with a greater degree of divergence. The real meaning is that for the comparison showing a greater numerical value, one or both samples are too large for our purpose, for that particular pair of populations with their spread and relative regularity of the frequency distributions and their difference between the means. Smaller samples would have been sufficient to prove what we set out to discover, if our purpose was the determination of the probable mathematical significance of the difference between the means. More specifically, when

two values of significance obtained in two comparisons are above its accepted limit, but differ widely—say, one is 10, the other 50—the meaning is that the samples in the latter comparison are larger than necessary for the purpose of determining significance. It is evident, therefore, that figures of different magnitudes which express mathematical probability or significance, cannot consistently be employed for the purpose of expressing relative divergence.

A notable example of a substitution of such constants is furnished by the "coefficient of racial likeness" which is extensively used by some physical anthropologists as a measure of population divergence. Regarding its proper use Morant (1923, pp. 205-207) states as follows:

"It [the coefficient of racial likeness] is not a true measure of absolute divergence, and must not for a moment be considered as such, but nevertheless we shall speak of it, for convenience, as if it were an absolute measure of racial affinity. When it is said that a low coefficient between two races A and B indicates a closer relationship than a higher coefficient between, say, A and C, what is meant always is that it is more probable that A and B are random samples from the same population than that A and C are."

This is a lucid statement of the underlying idea. The coefficient of racial likeness is essentially an expression of probability and not a measure of divergence. Only as an expedient make-shift is it used as a measure of divergence. It has been extensively used as such by Morant and others. However, a make-shift should be used only when it does not lead to false conclusions; but the coefficient of racial likeness often does lead to absurdly inaccurate biological conclusions, as shown by Seltzer (1937). It would seem to be best to abandon altogether the use of this coefficient as a measure of divergence, and if it is still desired to employ it as an expression of probability, to change its misleading designation. (The coefficient of racial likeness is used primarily to combine two or more characters for the purpose of measuring divergence. For any single character the misleading results obtained by using as a measure of divergence a certain formula that fundamentally expresses probability, is discussed by me in another place (1938, pp. 279-282). The problem of measuring divergence for a multiplicity of characters I have considered in another paper (1939).)

Physical anthropologists of the school of the London Biometric Laboratory having become inured to the use of the coefficient of racial likeness—which, as stated, is essentially an expression of probability—as a measure of racial divergence, we find a similar substitution of constants employed in still another connection. In a later paper, Morant (1936, p. 32) states as follows: "Different characters will arrange the series in very different orders, and it is not clear, at first, why more importance should be attached to one of these orders than to any other . . . A grading of the characters in order of importance for the purpose in view can be obtained by noting the number of significant differences found for each in a particular set of comparisons." He then lists the percentage of times, of the total number of comparisons made, in which each one of a number of characters showed a "significant" value for a (alpha is the chief, compound factor in the formula for determining the coefficient of racial likeness; Morant postulates that if a is greater than 10, it shows significance).

Now, what do we understand by an "important" character? Obviously a character is important in distinguishing populations when it manifests a comparatively high degree of divergence, and the opposite is true of an unimportant character. With respect to populations of specific or lower rank, the degree of divergence it shows is the criterion by which the importance of a character may be judged. The importance of characters in such populations may be considered from two points of view.

First, often a character may be said to be important in the sense that it may be employed to divide a number of related populations of similar taxonomic rank, such as a number of races within a species or a number

of species within a genus, into two major groups. Its importance then constitutes a group divergence. Such a character will show a comparatively high divergence when a population of one group is compared with a population of the other group, and a relatively low divergence when a pair of populations within either group is compared, although even within the limits of each group it will generally show different degrees of divergence to a certain extent. In such species or genera, when the populations are divided into pairs in all possible combinations and the pairs compared, the character, in general, will appear important in approximately half the number of comparisons and unimportant in the remainder; providing the number of populations in one group approaches equality to that of the other group, as they often do.

Second, more generally, the terms "important" and "unimportant" as applied to any given character is applicable only in connection with a given pair of populations, and they have no universal application. A character that may be important, that is, manifest a comparatively high degree of divergence, with respect to one pair of populations may be unimportant with respect to another pair, and *vice versa*. This is often true of a series of closely related populations. In a species containing many races, or in a genus comprising many species, that are divisible by important group characters into primary, secondary and tertiary groups, certain other important characters may crop up independently in some of the groups, and the same important character may appear in groups that are otherwise not immediately related. Often important characters thus appear in such kaleidoscopic fashion that they cannot be used consistently for the major division of the species in a genus or the races in a species. In general, therefore, a character may be said to be important only with respect to the comparison of a particular pair of populations.

Morant's attempt to determine certain characters as of general importance is, therefore, rather irrelevant. However, this is beside the point. What I am after is to point out that here also Morant uses a test of probability to express what is fundamentally a divergence.

In comparative biological research, the essential thing we are after, in general, is to determine a difference or a divergence. This is true of both morphological and physiological comparisons, using the latter term in a broad sense to comprise all life processes including the complicated chain of events connected with the reproductive process. Whether we compare the morphology of pairs of related populations in taxonomic work, the comparative yield of milk for a given breed of cattle in feeding experiments, the perceptible effects of a particular drug on guinea pigs or human beings in pharmacological research as compared with controls, etc., we are trying to determine the precise divergence between two variable quantities or populations (in cases similar to the latter, between treated and untreated individuals, or between the same individuals before and after treatment). This is our prime object. A secondary consideration is the mathematical determination of the probable reliability of the samples from a study of which the data are drawn that form the basis of our conclusions.

This being so, it is remarkable that hitherto most attention has been directed to the secondary consideration, the determination of probability, while the primary object, the determination of an adequate measure of divergence, has been rather neglected. A measure of divergence that is universally employed is the difference between the means of the two sets of data compared, but this is evidently not always adequate. It is certainly altogether inadequate in taxonomic research. A fundamental defect of measures of divergence in taxonomic research based on such values as the mean, median, or mode, is that they represent denominate numbers which are altogether unlike, their absolute values differing widely, in pairs of populations that differ by widely unlike characters. Consequently, the figures expressing the measures of divergence for different pairs of populations,

when based on denominate numbers, are not fairly comparable. That the figures are not comparable for characters the measures of which are expressed in different units is self evident; but even when expressed in the same unit they are often not fairly comparable, if the characters are unlike. For instance, if the divergence of a pair of closely related populations of mice be expressed by the difference of the means of the tail length measurements, and that of a pair of populations of fishes by the same difference of the head length measurements in the same unit, the relative divergence of the two pairs may not be fairly comparable. Furthermore, measures of divergence based even on the same character expressed in the same unit, are not fairly comparable for different pairs of populations if the spread of their distributions differ widely. An ideal measure of divergence, one that could be used as a universal yardstick, should be an abstract number based on the degree of overlap, positive or negative, of the two frequency distributions, such as the measure employed by me (1938). That measure appears to be fairly adequate for taxonomic work. Whether that measure, or a modification of it, will be found applicable to research problems similar to the other two mentioned above, I am not prepared to discuss. I am here speaking chiefly from the point of view of the taxonomist. (In another paper considering the measure of divergence with respect to a multiplicity of characters, I (1939) concluded that a measure of divergence based on the principal character is fairly adequate, and that if the other characters are to be considered at all in its determination, they are to be afforded minor weights. In any further attempt at the combination of several characters, the figures for the different characters used should be such abstract numbers that measure the divergence of every one separately, rather than denominate numbers that express their absolute values.)

Current formulae that are generally suitable for the determination of the probable reliability of samples investigated in taxonomic research, have as their fundamental bases the difference between the means of the two samples and their probable or standard errors, the size of the samples, and their variability as expressed by the standard deviation. The practical use of this determination in taxonomic research is rather limited. Fisher (1936, p. 59) states succinctly the proper, general application of the determination of significance as follows:

"It will be seen that the test of significance does no more, and attempts no more, than to answer the straightforward question, 'Could these samples have been drawn at random from the same population?' It calculates a probability. If the probability is very small the answer is 'No.' If it is not so small as to reach the level of significance required, the answer is 'Yes, they could.' The answer never is 'Yes, they must have been.' "

To this I may add that in taxonomic practice, in the majority of cases, the actual arithmetical determination is rather unnecessary. If the two frequency distributions are fairly regular (that is, the frequencies in the successive classes diminish successively at both sides of the mode, even though the distribution be skewed) as they usually are when based, respectively, on homogenous material and the sampling is adequate; and further, if the modes are at different even though closely adjacent classes, as they usually are when the two populations represented by the distribution really differ and the degree of divergence is rather considerable, the arithmetical determination of probability will usually result in a "significant" figure. Most of the cases covered by Fisher's first contingency may then be judged for practical purposes by a mere inspection of the data arranged in the form of frequency distributions.

In regard to pairs of distributions falling under Fisher's second contingency, that is, distributions showing a divergence of relatively low magnitude, and the differences of which do not reach the level of mathematical significance as determined by the samples examined; this class of examples will no doubt include many in which the differences are biologically signifi-

cant. In nature, differences in taxonomic characters between pairs of populations form a gradual series from small to large values, with virtually all possible intermediate values. (The series may be visualized as represented by a straight line of the equation, $mx-y=0$.) Small values near the extreme of the series must have a biological significance, although mathematically their significance appears doubtful. For such populations the arithmetical determination of probability is of no practical value by itself, because it fails to give a definite answer to the question in which we are interested, namely, is the difference real, even though small, or is it due to the vicissitudes of sampling? The mathematical answer to this question, to adapt Fisher's style in the preceding citation, virtually is "no" or "yes," which is no direct answer at all. When the test for significance results in a low numerical value, lower than the accepted limit, it may mean either one of two things: (1) The difference is not real. (2) The difference is real but its magnitude is such that the samples are not large enough to prove its reality mathematically. Larger samples are necessary for a mathematical test of significance. The meaning of too low a figure then may be similar but opposite to what was noted above that too high a figure for significance shows that the samples are too large; but when the figure expressing significance is high the answer is direct and positive, and when it is too low the answer is indirect and limited.

While some of the small but real differences that are not too extreme will show mathematical significance when the size of the samples are greatly increased—and theoretically any real difference, no matter how small, will show significance by taking samples that become infinite as the differences, in a series of pairs of populations compared, approach zero—in actual practice the size of the samples necessarily must be more or less limited. In workaday biological practice, therefore, it can hardly be doubted that small differences of biological significance will appear mathematically insignificant. In passing, it may be mentioned that instances may occur in which it would be impossible to obtain very large samples. Supposing we compare two populations of which the actual number of living individuals is very limited, and find a small difference which, based on the entire number of living individuals, does not show any mathematical significance. That does not mean that such a small difference does not have a biological significance. In general, in cases coming under Fisher's second contingency, our conclusions must be based on the biological evidence rather than on mathematical deduction.

From the standpoint of the comparative practical unimportance of the determination of probability in taxonomic research, it has received an undue share of attention from certain biologists whose work is essentially taxonomic, such as those investigations dealing with population or "racial" differences in various groups of living things. In general, this is also true of some statistical constants now in use in taxonomic work as noted below. From the point of view of the taxonomist at least, a great deal of what is being done along this line may be said to represent mathematical, rather than biologic research, employing biological data for the purpose of solving mathematical problems or formulating mathematical propositions. Of course, mathematics represents one of the important disciplines in the sum total of human culture, and there can be no objection for workers who are interested in mathematical research to illustrate their problems and propositions by the use of biological data, if they wish to do so. But it should be remembered that a great part of such research is of little importance in solving taxonomic problems. In taxonomic problems, what we are greatly interested in is to determine divergence as precisely as possible, while the determination of probability is of secondary importance. The thing to be regretted is not so much that a great deal of attention is being paid the latter, but that it is apparently done at the expense of the former. A few examples of published reports will illustrate this idea.

An outstanding, valuable and well known taxonomic work to which reference is often made in biological discussions is that by Crampton (1916, 1925, 1932) dealing with the terrestrial gastropods of the genus *Partula*. The investigation forming the basis of Crampton's reports is unusual as compared with taxonomic studies in general, by the number of specimens examined and the detail with which they were examined. The information furnished by Crampton makes it evident that *Partula* is at the present time level in an early and active stage of speciation, and as such, its detailed study is of special importance for an understanding of the process of evolution.

The study of speciation or raciation in *Partula* does not lend itself altogether readily to statistical treatment, because some of the important distinguishing characters are rather of a qualitative nature and are not readily expressible in terms of definite figures, although it is not altogether impossible to do so. One important character, the direction of the spiral of the shell, dextral or sinistral, can be expressed in terms of exact figures. In some species it is always either sinistral or dextral; while in other species, or populations of lesser rank, the direction of the spiral varies with the individual. In the latter, Crampton very helpfully gives the precise numbers or proportion of the sinistral and dextral individuals in his samples. For certain other characters that are measurable with more or less precision, Crampton furnishes a wealth of statistical data and constants in tabular form.

However, the tables published by Crampton furnish only a part of the information that his data were in position to furnish, and that not the most important part. For each character he generally gives the range of variation, the mean, the standard deviation and the error of the last two figures. For some species he also gives their coefficient of variation. Now, in the distinction of the species, or the different populations within a given species, and in the interpretation of the relationship of the various populations, of specific or lower rank, of what material difference, in general, is a knowledge that the standard deviation, or the coefficient of variation, in one is larger than in another population? Also, at their best such data are only approximate, and of what material difference, in general, is a knowledge of the small value of the error of the mean for the foregoing purposes? These figures are interesting, but they are largely of academic interest. Of course, there can be no objection if an author wishes to furnish such figures. What is regretable is that more pertinent information is omitted. From a taxonomic point of view we are intensely interested in how far or to what degree the different populations diverge with respect to the various characters. For that purpose we are presented only with the ranges and the means of the various characters, and these are altogether inadequate. To determine the precise extent of divergence, by some such method employed by me (1938), frequency distributions for the different characters for the separate populations are needed and these are omitted for the characters based on measurements.

For three characters Crampton does give frequency distributions, namely, the direction of the spiral of the shell, the degree of tooth development, and the color pattern. Crampton's presentation of the data for the last two characters is especially interesting, because they are rather qualitative in their nature. As such, their determination in terms of definite figures is only approximate and dependent to some extent on a subjective estimate. Qualitative characters are generally described by authors in adjectival words or phrases that necessarily must be indefinite to a certain extent, and not in terms of definite figures. Crampton shows that such characters also can be expressed, approximately, in the form of frequency distributions. Similarly qualitative characters in other groups as well may be expressed in figures, and although such figures necessarily must be only approximate at their best, they should yet prove to be of importance in

determining divergence between closely related populations. An interesting example of this kind is furnished also by the work of Sumner on mice of the genus *Peromyscus*. Characters based on color differences, in general, are qualitative, yet Sumner (1929) has found it possible to express quantitatively seven such characters, presenting his results for two of them in the form of histograms (p. 111), and giving the averages for the other five.

I have used Crampton's reports as an example because they constitute a work of unusual value and interest as compared with the ordinary run of taxonomic papers, but the foregoing statements apply to many other published papers in which taxonomists employ statistical methods. I may here cite three recent papers in my own specialty, in fishes, that happened to come to my attention, namely, by Schultz (1937), by Matsubara (1938) and by Storey (1938). These papers are much less extensive in scope than Crampton's reports, in that they deal with much fewer populations. They also differ more or less in the manner of the statistical presentation of the data; but they illustrate in different ways some of the points raised above.

Schultz compares the Pacific with the Atlantic population of the capelin. He compares a larger number of characters than usual in such cases, but for each character he publishes only the range of variation and the mean with its error. These figures are altogether inadequate for determining the precise divergence between the two populations, the thing in which Schultz as well as other taxonomists are chiefly interested. Had Schultz's data been published in the form of frequency distributions, they would constitute a valuable example showing the differing degrees of intergradation of the several characters in two closely related populations that differ by more than one character, in addition to forming a basis for the determination of the precise divergence between the two populations. (Schultz's method of combining several characters for the purpose of determining divergence I consider in another paper (1939).)

Matsubara, working with Japanese lizardfishes, does not employ statistical formulae or constants and does not calculate probabilities. Nevertheless, his method is essentially statistical in its nature, as it properly should be in a problem such as the author was confronted with. However, the data for the variability of the characters that are employed in comparing and distinguishing his populations (which happen to be of specific rank), are presented in graphic form and are not altogether suitable for the purpose of calculating the extent of divergence in terms of precise figures. Of course, the frequency distributions of the several characters may be approximately determined from the graphs, but it is very difficult or impossible to get the exact figures. For the precise determination of divergence, it is important to have the actual frequency distributions obtained during the investigation.

Storey, reporting on an investigation of the Atlantic populations of *Harengula*, also presents her data in graphic form and the same remarks apply to hers as well as to Matsubara's method of presentation. Furthermore, her data for characters having a continuous variation, namely, proportional measurements, are presented in the form of curves "smoothed by threes three times." "Smoothing" has the slight advantage of producing somewhat more regular curves which are rather more pleasing to the eye, but it has an important disadvantage in that the curves tend to mask heterogeneity in the material studied. That the material of *Harengula pensacolae*, for instance, possibly was heterogeneous is shown by her comparison (p. 35) of the specimens from Sanibel with those from other localities. Storey suggests that the differences in the measurements may be due to the different preservative used, formaldehyde instead of alcohol. This may be so to a certain extent, but part of the differences quite possibly represent a population divergence. The difference in the gill raker count of the Sanibel specimens would certainly seem to represent a population divergence. However, in clupeid species in general, the gill raker count differs greatly with the

size of the specimens and Storey does not appear to have segregated her data in sufficiently restricted size groups to reveal any possible intraspecific population differences in this character.

Detailed studies of other clupeid species have shown that they tend to diversification into distinct, statistically measureable, local populations of lesser rank, subspecies or races. It is highly probable that this is also true of the four Atlantic species distinguished by Storey. In view of the close approach or even general intergradation between these four species in the characters determined, it is quite possible that if such a detailed study be made, the relationship of the various populations will receive a modified interpretation than that obtained by the data available to the author. The size of the samples studied by the author were rather restricted (Storey, 1938, pp. 16-17), and in order to apply to them current statistical formulae, the grouping of the data adopted necessarily had to be comprehensive. A study of larger samples and of the same characters determined by the author, measurements, gill raker count and ventral scute count, with the data segregated by locality, and those of the measurements and gill raker count by smaller size groups, would possibly present a somewhat different picture of the relationship of the various populations, than that obtained by the grouping of the data as adopted by the author. It is evident that not only is it important—in order to determine precise divergence, distinguish properly the different populations and determine their relationship—to have frequency distribution tables published, but to subdivide the data where necessary by size, sometimes also by sex, and also by locality where heterogeneity is suspected.

Instead of presenting detailed frequency distribution tables, Storey gives derivatives of her data (table 3, pp. 16-17) in the form of certain constants, the most important of which are: the standard deviation, the mean, the difference between the means of the two populations compared and its standard error. These are not of much value in determining divergence as stated above. She also gives the relative deviate and the value of p, which express probabilities, as they are intended by the author to do, but are not suitable to determine divergence.

A paper based on the study of populations of flies that are of much interest in connection with some phases of the species problem, was very recently published by Mather & Dobzhansky (1939). It deals with the two well known "races" of *Drosophila pseudoobscura*, generally designated in the literature, following Lancefield's suggestion in his original report (1929) announcing their distinction, as race A and race B. The two populations occupy different but overlapping geographic ranges; they are also incompletely segregated ecologically (Dobzhansky 1937a, pp. 406-408).

The apparent principal character proving that the two populations are distinct is a physiological one and refers to the sterility of hybrid offspring when they are crossed. The sterility is partial, being confined to the males. Hybrid females are fertile, at least in part. A backcross of F_1 females to males of either parent population gives rise to both sterile and fertile males. Besides this principal character, Mather & Dobzhansky review and enumerate other, minor diverging physiological characters, and differences based on gene arrangement in the chromosones.

Morphological differences between the two populations that have so far been discovered show certain degrees of intergradation. The object of the paper mentioned is to deal with the morphological characters, and it takes up five such characters, namely, the number of teeth in both the proximal and distal sex comb on the leg of the male, the length and width of the wing and the length of the tibia. The former two are sex characters. The latter three characters were determined for males and females separately and they were found to differ by sex as well as by population. All five characters differed also according to minor populations or "strains" within each one of the two major populations.

Now, this problem is fundamentally taxonomic in its nature. Our concern is the determination of the relative rank of the taxonomic category in which the populations are to be placed. This determination, in its turn, must have as its basis a determination of the relative divergence of the populations. Given the known facts regarding the populations, let us consider this particular case from the taxonomist's viewpoint. This case may not be as remarkable as it appears. Every careful taxonomist of wide experience no doubt can cite similar instances in which distinct populations show relatively low and varying degrees of divergence with respect to morphological differences. Its apparent remarkableness rests on the partial sterility of the hybrid offspring correlated with a relatively low morphological divergence, and very likely is due to the fact that relative sterility and fertility of hybrids has been definitely determined only in a rather negligible number of very closely related populations.

In appraising the case under consideration taxonomically, it is well to consider the relative importance of physiological and morphological criteria in classification. There is no fundamental reason why the former should not be used for this purpose the same as the latter. Morphological criteria are generally used in taxonomy because they are determinable more readily and with greater precision. In the relatively few known cases in which a physiological character shows a greater divergence than any known morphological character, the former may be employed as the principal character in determining the taxonomic rank of the pair of populations compared.

In the case considered, the sterility criterion is evidently important, and we may confine ourselves to a consideration of this physiological criterion. The precise value of this criterion in classfication in general cannot be said to be as yet firmly established, and it cannot well be appraised, because it is known for relatively few populations as compared with their untold multitude. However, in general, it is evident that this criterion is not absolute, but, on the contrary, it is fully expressible in terms of degrees of magnitude only. Even in regard to the classical example of hybrid sterility, the mule, one now and then finds in the literature apparently authentic records of fertile individuals. Possibly, if sterility in the mule be investigated extensively and systematically, the percentage of fertile individuals may be found to be greater than such haphazard observations would seem to indicate. At any rate, judged by what we already know in regard to hybrid sterility in general, it is evident that this criterion shows all degrees of differences, from perfect or almost perfect sterility through different degrees of partial sterility to comparatively unhampered fertility, depending on the populations crossed.

There is an incomplete correlation between relative sterility and the relative degree of divergence of morphological characters. When an attempt is made to cross two closely related populations that have reached a sufficiently high degree of divergence, as determined by morphological criteria, to be generally regarded as species, one of several things may happen: (1) They may not be crossable. (2) They may produce zygotes showing various degrees of inviability, that is, they die at various stages of development, depending on the populations. (3) When viable offspring are produced, they may be sterile or show infertility of varying and rather pronounced degrees. On the other hand, when a cross is made between two populations the divergence of which is of such a rather low degree, as determined by morphological criteria, that they are generally regarded as of a taxonomic rank below that of species, fertile offspring generally seem to be produced. However, even with our present rather meager knowledge regarding sterility of hybrid offspring, it is evident that there are frequent exceptions to the above generalizations. A pair of closely related populations, which, judged by morphological criteria, are generally regarded as species sometimes, perhaps often, on being crossed give rise to a progeny that is fertile.

Since the magnitude of sterility is relative, being merely a matter of degree, it follows that if it be used as a criterion for grading populations into taxonomic categories, it would be necessary to draw arbitrary lines between the species and various categories of lower rank, the same as when morphological characters, especially quantitative ones, are used for that purpose. It is then necessary to devise a measure for expressing the degrees of sterility, and the most obvious measure that suggests itself is the relative percentage of sterile and fertile individuals in the hybrid progeny of the pair of populations compared. For instance, we may decide arbitrarily that if a pair of closely related populations on being crossed produce, on the average, a progeny 90% or more of the individuals of which are sterile, the populations are to be designated as species; they are to be designated as subspecies when the percentage of sterile individuals is 75 — 85%, other things being equal; they are to be designated as races when the same percentage is 60 — 70. These are the tentative arbitrary lines which I (1938) suggested to draw for morphological characters. For the sterility criterion even less data are extant than for morphological criteria, to enable us to draw the most pertinent arbitrary lines; but wherever drawn the lines evidently must be arbitrary. It may perhaps be found desirable to draw arbitrary lines for the sterility criterion that differ in numerical value from those employed for morphological criteria.

There being no absolute correlation between morphological divergence and relative sterility, the sterility criterion, if employed in taxonomy, evidently is to be used on a par with morphological criteria and coordinate with them. Whichever is the most divergent, it is to be used as the principal character for determining the taxonomic rank of the pair of populations compared. If any morphological character shows a divergence of specific magnitude the two populations are to be designated as species even though a cross between them produces offspring that are 100% fertile. Conversely, if the degree of sterility is greater than any morphological character that has been discovered, the former is the chief factor to be used in determining the taxonomic rank of the pair of populations compared.

Bearing the preceding propositions in mind, let us turn to the question of the taxonomic rank of the two major populations of *Drosophila pseudoobscura*. It is evident that for a pertinent decision of the question we need to know the degree of divergence of the various characters, morphological as well as physiological. For this purpose the paper by Mather & Dobzhansky furnishes only the averages of the morphological characters, which are entirely inadequate for measuring divergence, as discussed above. From a taxonomist's viewpoint, the data presented educe further questions. The authors very properly subdivide each one of the two major populations, races A and B, into minor populations which they designate as "strains." In the morphological characters the differences between the extreme strains within each race is nearly as great or greater than the differences between the major, composite populations. The question then is, when is a population to be designated as a "strain" and when is it to be designated as a "race"? More specifically, from the data presented it is evident that we have two (or more) "strains," one in each "race," that morphologically are approximately alike. What criterion then do they use for placing one strain in one race, and the other, morphologically similar strain, in the other race? Although I searched the paper for a definite statement in answer to this question, it could not be found. Apparently their basic criterion is sterility, for Dobzhansky (1937b, p. 285) states: ". . . the F_1 hybrid males from crosses between race A and race B are always sterile . . ." But other questions present themselves: On how extensive a body of data is the above quoted statement based? Especially, on how many crosses between different strains is it based? Is sterility of F_1 hybrid males absolute also between strains that are alike morphologically?, since we have seen that there is a rough correlation between morphological divergence and sterility. I did not deem it

necessary for the present purpose to enter into a complete analysis and review of the recorded investigations that have a bearing on an answer, if any, to the preceeding questions. Even should sterility of the hybrid male eventually prove not to be absolute between all the strains, it nevertheless seems evident from the above quoted statement that its degree, in general, is comparatively high.

Let us see now to what taxonomic conclusion we may come on the basis of the data that have been recorded.

A consideration of the condensed data presented by Mather & Dobzhansky makes it apparent that the morphological character showing the greatest divergence between the two major populations, taken in their entirety, refers to the number of teeth in the proximal sex comb of the male (with that based on the number in the distal comb a close second). While the precise degree of divergence between the two primary populations cannot be determined from the condensed data, it seems apparent that divergence is rather considerable but intergradation must also be pronounced. Arranging the averages in the authors' table 2 in their order of magnitude, five of the extreme minor populations, "strains," of race B have averages of 5.88, 5.92, 5.96 6.00 and 6.08 respectively, nearly the same as three extreme minor populations of race A with averages of 5.92, 5.96, and 6.00. The total number of minor populations compared are 20 of race B and 19 of race A; those having the averages nearly alike are, consequently, 25% of the total "strains" of B and 16% of A, or an average of about 20%. Had the authors given their data in the form of frequency distributions, the individuals comprised in the samples enumerated would apparently be seen to represent intergrades to a large extent. In addition, it seems apparent that a considerable number of individuals of the other minor populations, especially of those populations the averages of which are next in magnitude of those enumerated above, would also prove to be intergrades. It may be reasonably expected then that the two major populations taken in their entirety would show an intergradation of 25% or more in the morphological character of greatest divergence, according to the measure of divergence suggested by me (1938). According to the suggested arbitrary lines in the foregoing paper, this represents a divergence of racial or nearly subspecific magnitude. Judged by this criterion, on the basis of the incomplete data, the two populations perhaps are to be taxonomically designated as races, or they may possibly be near the borderline between the subspecies and the race.

The other important matter to consider is sterility. At stated, for the present it is difficult to form a judgment regarding the precise treatment of this criterion in taxonomy in general. We must also bear in mind the fact that in this particular case sterility is confined to one sex, the male. While by the use of morphological criteria species are sometimes based on characters of one sex, and properly so, the question remains whether the same course is to be followed with respect to the sterility criterion. Nevertheless, the fact of hybrid male sterility appears to be of tremendous biological significance, especially if we assume that it is 100%, or nearly so, as the above quotation from Dobzhansky would seem to indicate. Therefore, taking into consideration the sterility criterion, the rather considerable divergence in two morphological characters, and the other differences mentioned by Mather & Dobzhansky, the divergence would seem to be approximately of subspecific magnitude. A taxonomist's best and most reasonable judgment based on the extant incomplete evidence, would then seem to dictate the course, at least tentatively, of recognizing the two major populations as subspecies. According to a common taxonomic practice, it may be desirable to formally distinguish them by name, as, for instance, to designate them *Drosophila pseudoobscura*$_2$ and *Drosophila lancefieldi*$_2$, employing the notation for subspecies as suggested in my (1938) paper.

The course here suggested will give taxonomic expression to the relative divergence of some of the major populations of the genus *Drosophila* with

respect to morphological criteria and the sterility criterion. Some of the other major populations diverge in varying but pronounced degree morphologically, and when crossed produce inviable or altogether sterile offspring; they are, therefore, recognized as species. The two populations under consideration diverge morphologically in lesser degree and produce partly fertile offspring, and are consequently designated as subspecies.

The sterility criterion may be broadened to include the various graded results that occur when a cross is made between two populations. The grades are: incrossability, inviability of zygotes graded, in its turn, according to stage of development at which they die, offspring viable but sterile (even such populations may be only partly crossable), progeny partly fertile. It is interesting to note that in the one genus *Drosophila* there appears to be a correlation, at least partial, between hybridization results and morphology. Species that are readily separable morphologically appear to be either not crossable or to produce inviable zygotes. *D. melanogaster* and *D. simulans*, a cross of which results in viable but sterile offspring (Sturtevant 1929), also diverge morphologically in a lesser degree than some other species, so much so that they were not distinguished until comparatively recently (see Sturtevant, 1921, pp. 91-92). Finally, the two major populations of *D. pseudoobscura*$_1$ show a comparatively low morphological divergence correlated with partial fertility of hybrids.

The differences between *D. miranda* and *D. pseudoobscura*$_1$ (Dobzhansky, 1935) seem to be rather intermediate between that of the latter two pairs of populations mentioned. They produce viable but altogether sterile offspring the same as the cross between *D. simulans* and *D. melanogaster*. But morphologically the difference between them is evidently not greater than that between "races" A and B of *pseudoobscura*, whereas *D. simulans* and *D. melanogaster* are more greatly divergent morphologically, showing one apparently discontinuous difference, that relating to the structure of the male genitalia (compare Sturtevant, 1921, especially his figures 13-14, p. 34, with Dobzhansky, 1935).

Dobzhansky (1935) and Dobzhansky & Tan (1937) describe important differences in the chromosome structure between *D. miranda* and *D. pseudoobscura;* but, except in the peculiar distribution and number of sex chromosomes, such differences are evidently also nothing more than a matter of degree, since similar differences appear to exist not only between the two major populations of *pseudoobscura*, but also between the "strains" or minor populations (see Dobzhansky, 1937b, pp. 92-95). In general, hardly the surface has been scratched in elucidation of differences in chromosome structure between very closely related populations, and it cannot be used at present as a criterion in classification with any degree of assurance.

D. azteca and *D. athabascae* evidently constitute another example of two closely related populations of *Drosphila*, ". . . that are very similar externally but that produce sterile F_1 hybrids . . ." (Dobzhansky, 1937b, p. 113).

At least some of the major populations of the genus *Drosophila* have evidently reached at the present time level a fairly advanced stage of speciation, so that they are rather easily separable and are designated as species. Yet, even in this genus, when all major populations are considered, there is a certain gradual transition in degrees of divergence as determined by both morphological criteria, and physiological criteria based on hybridiztion. As for the minor, intraspecific populations, it may well be expected that various degrees of divergence will be discovered when the several species are subjected to such taxonomic analysis as was carried out by Mather & Dobzhansky on *D. pseudoobscura*$_1$. The same state of affairs very likely will be found to exist when we know more about the variability of some of the numerous poorly known species, or what are now recognized as species.

The taxonomic course suggested above indicates how a taxonomist would or should form a judgment in this particular case. A decision could be made with greater assurance, had the authors presented frequency distributions of at least the morphological character showing the greatest divergence. That would have given us a basis for a determination of the precise degrees of divergence between the minor populations and of the divergence of the two major populations by combining the data of each one of the two groups of minor populations, by some such method as was suggested by me (1938). Instead, the authors present only derivatives of their data in the form of means and squares of deviations for each character, and for a combination of characters, "a score," the latter by a method developed by Fisher. The main results of these figures is that they lead to other, final derivative figures expressing mathematical tests of significance of the differences between averages by standard methods. Now, just what are the values of the final figures in taxonomy? Do they convey any special meaning or ideas to the biologist, which would help him to come to a better decision, one made with greater assurance, in regard to the taxonomic status of the two populations, than that based on their morphological divergence and on the sterility criterion? The test of significance is interesting, but it is only of minor interest in taxonomy. In this instance especially, the definite determination of mathematical significance would seem to be of no more than academic interest. Arranging the averages given by the authors in their tables 2 and 4 in order of magnitude, one with some experience in applying current statistical formulae may see at a glance, without going through the actual arithmetical calculation, that the figures for the averages would result in values for the test of significance that would reach its accepted limits for the principal character. Any difference in the magnitude of the values beyond that limit does not have any special taxonomic meaning, as noted above. To an experienced and careful taxonomist, the figures for the averages, even condensed summaries though they are, speak much more eloquently, they constitute a much better basis for the formation of constructive decisions than the figures resulting from the test of significance. In sum, what the paper virtually accomplishes is to determine the figures for the test of significance. But this is only a secondary part of the problem. The main thing that we need to determine is the precise degree of divergence. This question is very inadequately answered.

While the greater part of the paper deals with the test of significance which is a matter of but minor interest in taxonomy, the subject of our primary concern, the determination of precise divergence to serve as a basis for forming a decision with assurance is incidentally considered, only insofar as the given averages form a very inadequate measure for such a determination. The valuable data determined by the authors in their investigation is not presented in such manner that the precise degree of divergence could be determined. Here then is an example of an investigation that bears the earmarks of care and reliability in its execution and carried out by reputable investigators, but the report of which fails to furnish the data in such form as will be of most help in deciding the question of our chief concern. To use a favorite expression of biological statisticians, the authors failed to extract all the information which their data were capable of furnishing, and the most important part is omitted. They fail to give even the ranges of variability of the different characters. This investigation is especially interesting in that, in a sense, it represents a study in experimental taxonomy. The groups of individuals, designated by the authors as "strains," upon which the data were determined, were bred in the laboratory from parents of known origin. It would be very interesting to compare divergence in such populations with that of the same populations as they occur wild; but the data as presented permit only an inadequate comparison.

I wish to emphasize here that I am not criticizing the paper as such. What I am after is to discuss its value in taxonomy, and the problem with which the paper deals is primarily a taxonomic one.

Physical anthropologists have been very assiduous in determining series of measurements of their material in many characters, as may be gathered by going through the volumes of *Biometrica,* for instance. Now, although they apparently have different standards than other systematists by which they determine the taxonomic rank of the populations studied by them, physical anthropology is nothing more than a highly specialized branch of taxonomy, dealing chiefly with one genus, *Homo,* and the same methods are applicable to this specialized branch as to taxonomy in general. In going through the published reports on physical anthropology in *Biometrica,* one finds that, omitting correlation studies, the figures on which the discussions and conclusions are based generally consist of the mean, the standard deviation, the coefficient of racial likeness, and their errors; sometimes the coefficient of variation is presented. However, what we are chiefly interested in, is the precise divergence between the populations, and those figures are not of much value in determining that, as stated.

The published reports that are used to illustrate the foregoing discussion, represent taxonomic investigations carried out by the use of appropriate modern methods, statistical methods. No fault can be found with the methods of investigation adopted by the authors. These are the methods that are to be recommended in taxonomic research. One or another of the investigations reported may be incomplete in one or another direction, but the methods adopted are correct as far as they go. Nevertheless they fail to determine the thing that is most important taxonomically, namely, the precise degree of divergence. This is generally true of published reports of taxonomic investigations in which modern statistical methods are employed. It seems that, in general, workers labor under the spell of statistical formulae developed through generations, and having to do chiefly with the theory of probability, something which is generally of but secondary importance in taxonomy.

What is to be especially regretted is that the class of reports under consideration, with frequent fortunate exceptions, omit the necessary data, complete frequency distribution tables, by which precise divergence could be determined by anybody who is interested. In general, in any investigation it is the data determined that are of primary importance. The manner of treatment of the data, their interpretation and the conclusions drawn from them, as given by the investigator, may not be the only possible ones. It is possible that some biologist at a later date, considering the subject from another point of view, may wish to treat the data in a different manner, and such treatment may even necessitate a different interpretation and lead to different conclusions. However, this would be impossible unless frequency distribution tables of the original data are presented. As it is, valuable data forming the quintessence of any investigation, representing a great deal of time and patient and concentrated effort, is thus practically lost to future workers. For instance, I (1938) employed a certain measure for determining precise divergence. This measure seems adequate in taxonomic work. However, some future worker may propose a measure that is more adequate, better expressive of the essential facts. But as most reports are presented now, it is not possible to determine divergence by the method employed by me, and will probably prove to be indeterminable by any future method that may be proposed. For the method mentioned at least, frequency distribution tables of the data are necessary, and this will likely prove to be so in any case.

In this connection a few words may not be amiss in regard to the economic aspect of the subject, having the editor's point of view in mind. Authors are sometimes confronted with the editor's desire to abbreviate manuscript reports by eliminating parts that appear not very essential. (It is unfortunate that this happens even with reports of outstanding merit.) In such cases, if elimination of some parts becomes necessary, and the question comes up whether to eliminate frequency distribution tables or graphic

representations of such tables, it is best to dispense with the latter. For, graphic representations merely constitute a device to "catch the eye" and clinch the author's conclusions, and it is usually possible to represent data graphically in more than one way, while the data on which the report is based are quintessential, as stated. If absolutely necessary, frequency distributions should be included even at the expense of parts of the discussion. In some cases, a report may be abbreviated even without urging from the editor, and yet be more informative than in the form in which it finally appears. For instance, some reports include tables giving detailed measurements of individual specimens, and yet they fail to include frequency distribution tables. But in mass data individual measurements are not of much consequence. (They may be of some interest in the case of very extreme or palpably aberrant specimens.) Such tables are rather superfluous and not likely to be carefully perused even by specialists directly concerned. In mass data what we are chiefly interested in, and what our conclusions are likely to be based on, is the frequency distribution of a population with respect to a given character under consideration.

It is fortunate for the cause of science that taxonomists are more and more abandoning the idea that taxonomy consists chiefly of the publication of local lists and catalogs, and descriptions of new species. This was all right for taxonomy in its pioneering stages, and if carefully and skillfully done, such papers still serve a useful purpose. However, gradually it is coming to be realized that the backbone of taxonomy is to be found in the careful and adequate comparison of related populations, whether they be of specific, subspecific or racial rank, by statistical methods, for the purpose of determining the intrapopulational variability, and interpopulationally, their precise degree of intergradation, or divergence, with each other. This forms a proper basis for an understanding of the relationship of groups of closely related populations. Our chief interest is to determine the precise divergence between pairs of closely related populations, or concomittantly their precise intergradation, these two values being complementary. The determination of probability is of but secondary importance in taxonomy, although most attention has hitherto been given to it. In issuing reports of taxonomic investigations in which statistical methods are employed, it is essential to include frequency distribution tables of the data, so that precise divergence, or intergradation, may be determined, by existing methods or by methods that may be discovered in the future.

Literature Cited.

Crampton, Henry E.

1916. Studies on the variation, distribution and evolution of the genus Partula. The species inhabiting Tahiti. *Publ. Carnegie Inst. Washington*, no. 228, 313 pp.

1925. The species of the Mariana Islands, Guam and Saipan. *Ibid.*, no. 228A, 116 pp.

1932. The species inhabiting Moorea. *Ibid.*, no. 410, 335 pp.

Dobzhansky, Theodosius

1935. Drosophila miranda, a new species. *Genetics*, vol. 20, pp. 377-391.

1937a. Genetic nature of species differences. *Amer. Nat.*, vol. 71, pp. 404-420.

1937b. Genetics and the origin of species. Columbia University Press, 364 pp.

Dobzhansky, T. & Tan, C. C.

1937. Studies on hybrid sterility. III. A comparison of the gene arrangement in two species, *Drosophila pseudoobscura* and *Drosophila miranda*. *Zeitsch. Induk. Abstam. Vererbl.*, bd. 72, pp. 88-114.

FISHER, R. A.
1936. "The coefficient of racial likeness" and the future of craniometry. *Jour. R. Anthrop. Inst. Great Britain and Ireland*, vol. 66, pp. 57-63.

GINSBURG, ISAAC
1938. Arithmetical definition of the species, subspecies and race concept, with a proposal for a modified nomenclature. *Zoologica*, vol. 23, pp. 253-286.
1939. The measure of population divergence and multiplicity of characters. *Jour. Washington Acad. Sci.*, vol. 29, pp. 317-330.

LANCEFIELD, D. E.
1929. A genetic study of crosses of two races or physiological species of *Drosophila obscura*. *Zeitsch. Induk. Abstam. Vererbl.* bd. 52, pp. 287-317.

MATHER, K. & DOBZHANSKY, T.
1939. Morphological differences between the "races" of Drosophila pseudo-obscura. *Amer. Nat.*, vol. 73, pp. 5-25.

MATSUBARA, KIYOMATSU
1938. A review of the lizard-fishes of the genus *Synodus* found in Japan. *Jour. Imp. Fish. Inst. Tokyo*, vol. 33, no. 1, pp. 1-36.

MORANT, G. A.
1923. A first study of the Tibetan skull. *Biometrica*, vol. 14, pp. 193-260.
1936. A contribution to the physical anthropology of the Swat and Hunza Valleys based on records collected by Sir Aurel Stein. *Jour. R. Anthrop. Inst. Great Britain and Ireland*, vol. 66, pp. 19-42.

SCHULTZ, LEONARD P.
1937. Redescription of the capelin Mallotus catervarius (Pennant) of the North Pacific. *Proc. U. S. Nat. Mus.*, vol. 85, pp. 13-20.

SELTZER, CARL C.
1937. A critique of the coefficient of racial likeness. *Amer. Jour. Phys. Anthrop.*, vol. 23, pp. 101-109

STOREY, MARGARET
1938. West Indian clupeid fishes of the genus *Harengula*. *Stanford Ichthy. Bull.*, vol. 1, pp. 3-56.

STURTEVANT, A. H.
1921. The North American species of Drosophila. Carnegie Inst., Washington, Publ. no. 301, 150 pp.
1929. The genetics of Drosophila simulans. *Ibid.*, no. 399, pp. 1-62.

SUMNER, FRANCIS B.
1929. The analysis of a concrete case of intergradation between two subspecies. *Proc. Nat. Acad. Sci. U. S.*, vol. 15, pp. 110-120, and 481-493.